DEVELOPMENT OF ASSESSMENT INSTRUMENTS HIGHER ORDER THINKING SKILLS ON SCIENCE SUBJECTS FOR STUDENT GRADE EIGHT JUNIOR HIGH SCHOOL

Khoiriah
Science Education Master Department, Lampung University, Indonesia
khoiriahspd74@gmail.com

Tri Jalmo
Dr., Science Education Master Department, Lampung University, Indonesia
jalmotri@yahoo.com

Abdurrahman
Dr., Science Education Master Department, Lampung University, Indonesia
abeunila@gmail.com

ABSTRACT
This study aims to produce an assessment instrument that meets the eligibility criteria as a higher order thinking skills (HOTS) assessment instrument. The research procedure adapted from the R & D model refers to Gall, et al. (2003). The design of the main field test study used pre-experimental type one-shot case study. Subjects involving 174 students of grade nine at Junior High School in Bandar Lampung at Lampung Province, Indonesia with random sampling technique. The eligibility criteria for HOTS assessment instruments were collected using theoretical validation sheet and empirical validity. Data analysis is done by descriptive statistic and anates program. The result of data analysis show theoretical validity covering material aspect 84,00% (valid); construction 93,35% (very valid); language is 87,13% (very valid) whereas empirical validity includes the validity coefficient of 0,70 (high); reliability 0,82 (very high); difficulty level 7,50% (difficult): 77,50% (moderate): 15,00% (easy); power difference 10,00% (good) and 90,00% (sufficient); functionality of deception 67,50% (very good) and 32,50% (good).
Keywords: development, assessment instruments, higher order thinking skills, theoretical validity, empirical validity.

INTRODUCTION

Teaching competence is a basic skill that professional educators must possess. Before the teacher learning process begins the agenda by designing learning activities. One of the important aspects that should be in the planning of learning activities is to set teaching objectives. Based on the plan and the purpose of teaching teachers to carry out learning activities. To provide an overview of the learning process and progress as well as the achievement of student learning, the teacher conducts assessment or assessment. Assessment is an integral part of the overall learning process (Santrock, 2014).

Assessment provides a variety of information that can be used as a basis for decision-making on students, curriculum, learning programs, school climate, and school policy (Uno & Koni, 2014). Assessment is often used by teachers in measurement and non-measurement processes to obtain data related to student characteristics (Popham, 1995). So the quality of the assessment instrument determines the accuracy of the teacher when improving the student's learning outcomes and results (Indrastoeti, 2012).

Quality assessment tools are instrumental in helping the learning process, identifying the strengths and weaknesses of learning, assessing the effectiveness of the learning strategies used, and providing data to assist students in making decisions to improve behavior and learning environments (Kusairi, 2012). The assessment approach now focuses more on getting students thinking, reasoning, and taking an active role in learning, not just looking at things that students remember or report or simply show students to perform calculations or be able to perform the procedures correctly (Mardapi & Andayani, 2012).

Changes in the focus of assessment standards have changed the way the world views the world of science. Science today is no longer only seen as a science but is also needed as something that can be used to survive (NRC, 1996). Therefore, in order to prepare students to face the challenges of life in the future, the objectives of the assessment of science learning lead to a more productive thinking that encompasses critical and creative thinking as well as developing higher order thinking skills (HOTS) (NRC, 2003).

HOTS development in learning can train students accustomed to thinking HOTS (Saido, et al., 2015). HOTS encourages students to apply knowledge and skills in new contexts, use new information and "manipulate" information to reach possible answers to new situations, and enable students to store information and implement real-world solutions (Brookhart, 2010; Heong, et al., 2011; Ramos, et al., 2013).

Taxonomy Bloom revised is fundamental to the development of thinking skills. Cognitive domain of remembering (C1), understanding (C2), and applying (C3) in Taxonomy Bloom revised including lower order thinking skills (LOTS) while HOTS covers the cognitive domain of analyzing (C4), evaluating (C5), and creating (C6) (Anderson & Krathwohl, 2001). However, the development of HOTS in science learning has not been optimal, more oriented teachers develop LOTS than HOTS (Depdiknas, 2008; Khoiriah, 2017). This situation has had an impact on the low achievement of science students in several countries around the world, as reflected by the results of the analysis of the achievement of science students of grade eight Junior High School as released by the 2015 TIMSS (Trends in International Mathematics and Science Study) mapping study showing 15 out of 49 participating countries are in a position below the international average score below 500 and Indonesia is ranked 36 (IEA, 2016).

The not yet optimal development of HOTS in science learning is a manifestation of the weak competence of science teachers to prepare HOTS assessment instruments. As the data shows most of the science-made science question items are cognitive landings of C2 (81,82%) and a small part including the C4, C5, C6 cognitive domain of 9,09% - 18,18% (Khoiriah, 2017). This data proves that science teachers have difficulties when composing HOTS assessment instruments.

The form of HOTS assessment instruments may be multiple choice items (Kubiszyn & Borich, 2013). HOTS assessment instrument preparation technique is almost the same as LOTS, only because students are tested in cognitive domain analyze, evaluate, and create then there must be component or stimulus that can be analyzed, evaluated and created (Devi & Widjajanto, 2011). The stimulus of the HOTS assessment instrument may take the form of a reading source, case, image, graph, photograph, formula, table, list of words or symbols, samples, films or sounds recorded (Brookhart, 2010). In addition, HOTS assessment instruments are non-algorithmic, complex, have many solutions, involving variations in decision making and interpretation, applying many criteria, and requiring much resolving (Resnick, 1987).

Based on the above problems, the authors have conducted research on the development of HOTS assessment instruments that meet the eligibility criteria as a HOTS assessment instrument based on theoretical validity and empirical validity.

METHOD

Research Stages

This research is a development research using research and development (R & D) procedure which refer to Gall, et al., (2003). This research adapts 7 out of 10 stages of R & D model Gall, et al., (2003): (1) research and information gathering, (2) planning, (3) initial product development, (4) expert team testing, 5) revision of initial product test results, (6) main field testing, (7) revision of main field test results, (8) operational field tests, (9) revisions of operational field test products, and (10) implementation and dissemination .

Research and information gathering stages include literature studies and field studies. The literature study was conducted to obtain the data as the theoretical foundation in strengthening the argumentation for the product of the development result while the field study aimed to obtain data related to the knowledge and experience of the teacher compile the HOTS assessment instrument. The planning stage covers the design of the HOTS assessment instrument. The testing phase of the expert team involves a team of expert science assessments and evaluation experts with the focus on theoretical validation testing. The main field testing stage focuses on the empirical validation test of the initial product.

Product Development Results

The product of development in this research is HOTS assessment instrument in the form of multiple-choice items on human circulation system for grade eight at Junior High School (Kubiszyn & Borich, 2013) students. The HOTS assessment toolkit consists of 20 items on HOTS package A and 20 items on HOTS package B with cognitive domain analyzing (C4), evaluating (C5), and creating (C6) (Anderson & Krathwohl, 2001; Brookhart, 2010; (Cognitive), which includes factual knowledge (K1), conceptual knowledge (K2), procedural knowledge (K3) and knowledge of metacognition (K4) (Anderson & Krathwohl, 2010).

Eligibility Criteria for HOTS Assessment Instrument Development Results

The quality of the HOTS assessment instrument's feasibility assessment is ensured through theoretical validity and empirical validity. Theoretical validity includes material aspects of "valid" categories, "valid" categorized construction aspects and "valid" categories of language aspects, while empirical validity includes the validity of the item with minimal "sufficient" interpretation, reliability with "high" interpretation, the difficulty level with the proportion of 15% easy: 80% moderate: 5% difficult, and distinguishing power with minimal interpretation "enough" and 80% of the deception are "good" (Nofiana, et al., 2016).

Data Collection Technique and Data Analysis

The data of this research include qualitative and quantitative data. Qualitative data obtained from the results of expert team test assessment while quantitative data from the initial field test results of the initial product.

Qualitative data collection is done by using theoretical validation sheet. The theoretical validity sheet is a list of likert-scale questions so the team of experts only gives checklist (√) marks on the "1 (invalid)", "2 (less valid)", "3 (simply valid)", "4 (valid)" , and "5 (very valid)" in the available column according to the assessment then gives the final conclusion by circling one of the options of LD (feasible to use), LDP (feasible to use with improvement), or TDL (not worth using) (Table 1). Based on the choice of expert team assessment is done percentage calculation, then the data interpreted using validation criteria according to Ratumanan, et al. (2009) (Table 2).

Expert team assessment results are used as a basis for revising the initial product. If based on the calculation using the content validity ratio (CVR) formula shows the validity number is less than the minimum limit of 0,60 then the initial product must be revised again. After the initial product is revised then the expert test returns to obtain a validity price of at least 0,60 or 2 expert validators interpret "LD (feasible use)" (Ratumanan, et al., 2009).

Table 1. Theoretical Validity Sheet HOTS Assessment Instrument Result of Development

Rated Aspect	Indicator	Assessment Scale/Item				
		1	2	3	4	5
Material	1. Conformity of item with the indicator.					
	2. Linkage distractor with the subject matter.					
	3. Conformity of the item with the type of school or grade level.					
	4. The suitability of the item with the HOTS cognitive domain tested.					
	5. There is only one answer key.					
	6. The conclusion of material aspect validation.	LD		LDP		TDL
Construction	1. The subject matter is formulated clearly and firmly.					
	2. Choice of answers is homogeneous and logical.					
	3. The subject matter does not provide a clue to the key answer.					
	4. The subject matter does not contain double negative statements.					
	5. The length of the formula is relatively the same answer.					
	6. The choice of numerical answers is arranged chronologically.					
	7. Pictures, graphs, tables, diagrams, and discourses on the matter are clear and functional.					
	8. The item does not depend on the previous answer.					
	9. The instructions for working on the problem clearly.					
	10. There is a scoring guide.					
	11. Each subject has a stem that students will use to think HOTS.					
	12. Conclusion of construction aspect validation.	LD		LDP		TDL
Language	1. Grammar and spelling in accordance with Indonesian rules.					
	2. The language used in accordance with the level of student development.					
	3. The conclusion of language validation.	LD		LDP		TDL

Table 2. Criteria for Achieving CVR Validation

Percentage	Criteria
21,00 – 36,00	Invalid (IV)
37,00 – 52,00	Less Valid (LV)
53,00 – 68,00	Simply Valid (SV)
69,00 – 84,00	Valid (V)
85,00 – 100,00	Very Valid (VV)

(Source: Ratumanan, et al.,2009)

Quantitative data collection aims to obtain initial empirical product validity data measured through analysis of student response responses in terms of item validity, reliability, difficulty level, distinguishing power, and function permit. The design of the study on the main field test using the pre-experimental designs type one-shot case study that is by giving the initial product in one group of 174 students of grade nine at Junior High School in Bandar Lampung at Lampung Province, Indonesia, as shown in Table 3 (Sugiyono, 2011).

Table 3. Research Design of Main Field Test Stage

Treatment	Results
X	O

Description: X = Treatment; O = Result

The response of the students' answers on the main field test was analyzed using an anates program. The results of data analysis serve as the basis for assessment and revision of the initial product. If the result does not meet the criteria of empirical validity eligibility as HOTS assessment instrument then the revised product is then continued to conduct the 2nd major field test, and so on until the initial product is produced that meets the criteria of empirical validity as an instrument of HOTS assessment. If the main field test results have met the criteria of empirical validity eligibility as a HOTS assessment instrument then the initial product hereinafter referred to as the final product of the development result.

RESULTS

Expert Team Test Results

The recapitulation result of the theoretical validity interpretation of the material aspects of the initial product according to the validation criteria (Table 4).

Table 4. Interpretation of Theoretical Validity Aspects of Initial Product Material

Number Item	Amount Question	Validation Accomplishment (%)	Criteria
Package A			
1, 2, 5, 6, 7, 8	6	88,00 – 96, 00	VV
3, 4, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20	14	72,00 – 84,00	V
Package B			
5, 6, 7, 8, 11, 12, 13	7	88,00 – 92,00	VV
1, 2, 3, 4, 9, 10, 14, 15, 16, 17, 18, 19, 20	13	72,00 – 84,00	V
Average Validation Accomplishment (%)		**84,00**	**V**

Description: VV = Very Valid; V = Valid

Based on Table 4 when referring to validation achievement criteria according to Ratumanan, et al. (2009), then the theoretical validity of the initial product material aspect can be declared "valid".

The result of recapitulation of the theoretical validity interpretation of the construction aspects of the initial product according to the validation criteria (Table 5).

Table 5. Results of Interpretation of Theoretical Validity of Early Product Construction Aspects

Number Item	Amount Question	Validation Accomplishment (%)	Criteria
Package A			
1 – 20	20	85,00 – 99,09	VV
Package B			
1 – 20	20	84,54 – 99,00	VV
Average Validation Accomplishment (%)		**93, 35**	**VV**

Description: VV = Very Valid

Based on Table 5 when referring to validation achievement criteria according to Ratumanan, et al. (2009), then the theoretical validity of the initial product construction aspect can be stated "very valid".

The recapitulation result of the theoretical validity interpretation of the language aspects of the initial product according to the validation criteria (Table 6).

Table 6. Interpretation of Theoretical Validity Aspects of Early Product Languages

Number Item	Amount Question	Validation Accomplishment (%)	Criteria
Package A			
1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 18, 19, 20	19	85,00 – 95,00	VV
17	1	80,00	V
Package B			
1, 3, 4, 5, 7, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20	17	85,00 – 100,00	VV
2, 6, 8	3	75,00 – 80,00	V
Average Validation Accomplishment (%)		**87,13**	**VV**

Description: VV = Very Valid; V = Valid

Based on Table 6 when referring to validation achievement criteria according to Ratumanan, et al. (2009), then the theoretical validity of the initial product language aspect can be expressed "very valid".

Main Field Test Results

The results of quantitative data quantitative analysis, reliability, difficulty, distinguishing, and derivative functions of the initial product can be seen in Table 7, Table 8, Table 9, and Table 10.

Table 7. Results of Recapitulation of Data Validity and Reliability of Initial Product Item

Package Item	Amount Question	Validity Item		Reliability Item	
		Coefficient of Validity	Validity Criteria	Reliability Coefficient	Criteria for Reliability
A and B	40	0,70	High	0,82	Very high

Based on Table 7 it can be stated that the initial product has "high" validity criteria and "very high" reliability.

Table 8. Result of Data Recapitulation of Level of Problem of Initial Product Item

No.	Criteria for Tribune	Number Item	Amount Question	(%)
1.	0,00 s.d 0,25 (difficult)	**Package** A = 8, 13	3	7,50
		Package B = 8		
2.	0,26 s.d 0,75 (Moderate)	**Package** A = 1, 3, 4, 5, 6, 7, 9, 11, 12, 14, 15, 16, 17, 18, 19	31	77,50
		Package B = 1, 3, 4, 5, 6, 7, 9, 11, 12, 13, 14, 15, 16, 18, 19, 20		
3.	0,76 s.d 1,00 (Mudah)	**Package** A = 2, 10, 20	6	15,00
		Package B = 2, 10, 17		

Based on Table 8 it can be seen that more than half of the initial product has a difficulty level of " moderate " criteria and a small number of "difficult" and "easy" categories.

Table 9. Results of Recapitulation of Differential Power Data of Initial Product Item

No.	Criteria Distinct Power	Number Item	Amount Question	(%)
1.	≥ 0,50 (Good)	**Package** A = 6	4	10,00
		Package B = 4, 11, 18		
2.	0,20 s.d 0,49 (Enough)	**Package** A = 1, 2, 3, 4, 5, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20	36	90,00
		Package B = 1, 2, 3, 5, 6, 7, 8, 9, 10, 12, 13, 14, 15, 16, 17, 19, 20		

Based on Table 9 it is revealed that almost all initial product differentiation power is categorized as "enough" and only a few criteria are "good".

Table 10. Result Recapitulation of Data Deception Function Item of Initial Product

No.	Criteria Deception Function	Number Item	Amount Question	(%)
1.	Very good	**Package** A = 1, 4, 5, 6, 7, 10, 11, 13, 14, 15, 16, 17, 18, 19, 20	27	67,50
		Package B = 2, 5, 6, 7, 8, 9, 10, 11, 14, 15, 16, 19		
2.	Good	**Package** A = 2, 3, 8, 9, 12	13	32,50
		Package B = 1, 3, 4, 12, 13, 17, 18, 20		

Table 10 shows more than half of the items in the original product having the cluster functioning "very good" and only a small portion functioning "good".

DISCUSSION

This development research aims to produce an assessment instrument that meets the criteria as a HOTS assessment instrument. The feasibility of the HOTS assessment instrument is guaranteed by theoretical validity and empirical validity. This is confirmed by Kartowagiran, et al. (1999) the quality of HOTS assessment instruments can be seen in terms of theoretical and empirical. As the results of the research Nofiana et al. (2016) describe the assessment instrument as a HOTS assessment instrument if it has the validity criteria of the "valid" material aspect, the "valid" construction aspect and the "valid" language aspect and the empirical validity with the validity criteria of item "intermediate", reliability of "high" interpretation, 15% easy proportion 15% difficulty: 80% moderate: 5% difficult, and distinguishing power interpreted to be minimal "enough" and 80% of deception are "good".

The material aspect test by the expert team was conducted to determine the material feasibility of the assessment instrument (Kartowagiran, 2011b). Material validity includes the alignment of instrument materials with learning indicators (Kartowagiran, 2012). The validity of the instrument material is related to the suitability and competence representation that must be achieved by the students so that the instrument improvement process must be done (Hendryadi, 2014). This is in line with Matondang (2009) to examine the question or revelation of the instrument needs to be done in order to obtain adequate quality question device.

The validity of the instrument material is determined by the justification of the experts (Istiyono, et al., 2014; Santyasa, 2005; Listyani & Hidayati, 2010; Muljono, 2007; Raisanen, et al., 2016). In this regard, the expert team has specifically provided suggestions for improvement, and then performed the process of improvement until the average achievement of theoretical validity of material material aspects of 84,00% then theoretically the material aspects of the HOTS assessment instrument development results categorized as "valid" (Ratumanan , et al., 2009).

The material aspect of the development result instrument is categorized as "valid" to reveal the set matter that has been compiled in accordance with the subject matter that must be achieved by the students at its level. As Arikunto (2011); Ramadhani, et al. (2014) describes the validity of instrumental material is said to be "valid" is able to measure the specific objectives of learning and in accordance with indicators in basic competencies. The "valid" category of the HOTS development assessment instrument proves that the instrument meets the eligibility criteria as a HOTS assessment instrument (Nofiana, et al., 2016).

In addition to meeting the eligibility criteria of material validity, assessment instruments must also be tested in construction (Sugiyono, 2011). The validity level of instrument construction can be known by reviewing the composition of the item. Testing the quality of theoretical validity of the construction aspects of the assessment instrument can be done by referring to the writing of the question and through the assessment of expert teams (Amarila, et al., 2014; Mardapi, 2003; Kartowagiran, 2012).

Based on the results of the assessment of the expert team of HOTS assessment instrument construction aspects of the development results show the average validation achievement of 93,35%. This data proves the construction aspects of categorized instruments as "very valid" (Ratumanan, et al., 2009) and reveals that the HOTS development assessment instrument has exceeded the eligibility criteria for HOTS assessment instruments (Nofiana, et al., 2016).

When composing an assessment instrument, language validity should also be given attention. This is in harmony with Rahayu, et al. (2014) that language validity aims to determine the accuracy of language usage in instrument devices. Moreover, if assessment instruments require a high level of reasoning, such as HOTS-oriented assessment instruments. As Sunarti & Rahmawati (2014) describes the language used in the assessment instrument must be communicative, in accordance with the Indonesian grammar and spelling rules, as well as the level of student development.

With regard to the language aspects of the HOTS assessment instrument, the result of the development of a team of experts has provided suggestions for improvement so that students can easily understand the questions and the answers asked to be clear. Based on the evaluation result of the expert team on the instrument language aspect shows the average validation achievement of 87,13% means that the instrument is categorized as "very valid" (Ratumanan, et al., 2009). Thus the HOTS assessment instrument of development result is stated to have eligibility criteria as a HOTS assessment instrument (Nofiana, et al., 2016).

After the material aspects, construction, and instrument language of development result meet the eligibility criteria as HOTS assessment instrument, then the instrument is tested on 174 students of grade nine at Junior

High School in Bandar Lampung at Lampung Province, Indonesia to know the quality of the empirical validity of the instrument. As Kartowagiran, et al. (1999); Lababa (2008); Kartowagiran (2011a) confirms to produce high-quality test questions, so in addition to tested theoretically also need to be tested empirically.

The question device is said to be of good quality if it has empirical validity including the coefficient of validity and reliability, distinguishing power, difficulty level, and distribution of choice of answer or deceiver (Ramadhani, et al., 2014; Budiman & Jailani, 2015). Empirical validity is obtained through test results (Matondang, 2009). Testing the problem is an effort to know the quality of questions based on the response of the test participants (Kartowagiran, 2012).

Based on the results of responses responses students obtained the value of the coefficient of validity and reliability of HOTS assessment instrument results of 0,70 and 0,82 respectively. Referring to the criteria of validity and reliability coefficients, the instruments are categorized as "high" and have "very high" reliability (Arikunto, 2011). Interpretation of the validity and reliability of the HOTS assessment instrument of the development results proved to exceed the minimum threshold of the eligibility criteria as HOTS assessment instruments. As the results of Nofiana's research, et al. (2016) asserted that the assessment instruments meet the eligibility criteria as HOTS assessment instruments if they have a validity of at least " sufficient " interpretation and minimum "high" interpretation reliability.

The above facts are similar to Mardapi (2003) that the instrument is said to have a reliability index or "good" reliability if the reliability coefficient index is at least 0,70. Furthermore, Dwipayani (2013) also describes the matter of the declared device of good quality if it has a high validity and reliability index.

The high level of validity and reliability of the problem due to HOTS assessment of development results has been through the justification process of the expert team. This fact proves the process of improving the HOTS assessment instrument of development results in accordance with the direction and suggestions of improvement from the expert team (Raisanen, et al., 2016; Amarila, et al., 2014).

In addition to the validity and reliability coefficients, the quality of the assessment instrument is also supported by the level of difficulty, distinguishing factor, and effectiveness of the deception function (Kartowagiran, 2012). Preparation of the questioning device should consider the level of difficulty in order that the results achieved can illustrate actual student achievement (Dwipayani, 2013). The analysis of the difficulty level is important because it can examine problems that are easy, moderate, and difficult to balance the proportion of easy, moderate, and difficult categorical problems in assessment instruments (Wardany, et al., 2015).

Based on the results of the response analysis of students can be known the level of difficulty of HOTS assessment instrument development results show from total 40 items tested there are 3 items (7,50%) range of 0,00 to 0,25 classified as "difficult", 31 items (77,50%) ranged from 0,26 to 0,75 with " moderate " criteria, and 6 item points (15%) ranged from 0,76 to 1,00 in "easy" categories. The proportion of difficulty level of HOTS assessment instrument of this development result proved in harmony with the results of Nofiana research, et al. (2016) stating that the assessment instrument is eligible as a HOTS assessment instrument if it has a 15% easy proportion of difficulty: 80% moderate: 5% difficult.

Furthermore, the difficulty level of the HOTS assessment instrument of development result is including the assessment instrument with the "good" item quality. As Mardapi (2003) asserts that the "good" item has an index of difficulty levels of between 0,30 and 0,80. Even Arikunto (2011) further emphasized that the item "good" is a matter that is not too easy and not too difficult.

Analyzing the feasibility of differentiating power in the assessment instrument needs to be done. This is because analyzing the differentiator is an activity to examine the items to know the ability of students in solving the problem (Uno & Koni, 2014). Further Rofiah, et al. (2013); Suwarto (2011) explained that analyzing the differentiation of the problem means measuring the ability of the item to distinguish between high and low group students based on certain criteria.

Furthermore, based on the results of data analysis of responses of students revealed also the difference in power index HOTS assessment results of development is 4 items (10%) has a different power coefficient ≥ 0,50 "good" criteria and 36 items (90%) coefficient power difference between 0,20 – 0,49 is "sufficient". Or the average of the differentiating power coefficient of 0,43 with "sufficient" criteria means that the items in the HOTS assessment instrument of development outcome are acceptable and stated to have satisfied the eligibility criteria as HOTS assessment instruments (Nofiana, et al., 2016). This is as supported by Yusuf (2015) and Mardapi

(2003) that the items that have differentiating power are "sufficient" to mean that the item is of good quality and can be used at a later stage.

The quality of multiple-choice items is tested when it comes to the quantitative analysis of the effectiveness of the deception function (Mardapi, 2003). Referring the results of the analysis of the fool function revealed the HOTS assessment instrument of the development result there are 27 items (67,50%) have the duties function "very good" and 13 item (32,50%) function "good". This data reveals the distribution of choice of answers in the question device has performed its function well as a deception and deserves to be declared as HOTS assessment instrument. As the results of Nofiana's research, et al. (2016) report assessment instruments are HOTS-certified if 80% of deception are "good". Further Rofiah, et al. (2013) describes a good performer when selected by 5% of test takers.

CONCLUSION

Based on the result of research and development it is concluded that the HOTS assessment instrument of development result has fulfilled the eligibility criteria as HOTS assessment instrument, that is the validity of the theoretical aspects of the material "valid" category, categorized construction is "very valid", and the language aspect is "very valid" while empirical validity has coefficient validity about 0,70 is "high", reliability is 0,82 "very high", difficulty level is 7,5% difficult: 77,5% moderate: 15% easy, distinguishing power equal to 0,43 classified criteria "enough", the effectiveness of the deception of 67,50% works "very good" and 32,50% works "good".

REFERENCES

Amarila, Raula Samsul., Habibah, Noor Aini., & Widiyatmoko, Arif. (2014). Pengembangan Alat Evaluasi Kemampuan Berpikir Kritis Siswa pada Pembelajaran IPA Terpadu Model Webbed Tema Lingkungan. *Jurnal Pendidikan Universitas Negeri Semarang.* 3(2),563-569.

Anderson, L.W., & Krathwohl, D. R. (2001). *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives.* Allyn dan Bacon. Addison Wesley Longman. New York. (2010). *Kerangka Landasan untuk Pembelajaran, Pengajaran, dan Asesmen.* Pustaka Pelajar. Yogyakarta.

Arikunto, Suharsimi. (2011). *Dasar-Dasar Evaluasi Pendidikan*. Bumi Aksara. Jakarta.

Budiman, Agus., & Jailani. (2015). Developing An Assessment Instrument of Higher Order Thinking Skills (HOTS) in Mathematics for Junior High School Grade VIII Semester 1. *Proceeding of International Conference on Research, Implementation, and Education of Mathematics and Science.*

Brookhart, Susan M. (2010). *How to Assess Higher Order Thinking Skills in Your Classroom.* Alexandria, Virginia. The USA.

Clark, D. (2010). *Bloom's Taxonomy of Learning Domains: The Three Types of Learning.* Retrieved from http://www.nwlink.com/~donclark/hrd/bloom.html.

Depdiknas. (2008). *Strategi Pembelajaran Matematika dan Ilmu Pengetahuan Alam.* Direktorat Jendral Peningkatan Mutu Pendidik dan Tenaga Kependidikan. Departemen Pendidikan Nasional. Jakarta

Devi, Poppy Kamalia., & Widjajanto T, Erly Tjahja. (2011). *Instrumen Penilaian Hasil Belajar IPA High Order Thinking*. P4TKIPA. Bandung.

Dwipayani, Anak Agung Sri. (2013). Analisis Validitas dan Reliabilitas Butir Soal Ulangan Akhir Semester Bidang Studi Bahasa Indonesia Kelas X SMA Terhadap Pencapaian Kompetensi. *Jurnal Pendidikan.*1-18.

Gall, P. M., Gall, J.P., & Borg, W.R. (2003). *Educational Research An Introduction. Seventh Edition.* University of Oregon, Boston. New York.

Hendryadi. (2014). Validitas Isi. *Teori Online Personal Paper.*1(6), 1-5.

Heong, Y.M., Othman, W.D., Md Yunos, J., Kiong, T.T., Hassan, R., & Mohamad, M. M. (2011). The Level of Marzano Higher Order Thinking Skills Among Technical Education Students. *International Journal of Social and Humanity.*

Indrastoeti, Jenny. 2012. *Pengembangan Asesmen Pembelajaran*. UPT UNS Press. Surakarta.

International Organization for Evaluation of Educational Achievement(IEA). (2016). *Annual Report.* Retrieved 26 April, 2017 from www.iea.org.

Istiyono, Edi., Mardapi, Djemari., & Suparno. (2014). Pengembangan Tes Kemampuan Berpikir Tingkat Tinggi Fisika (PysTHOTS) Peserta didik SMA. *Jurnal Penelitian dan Evaluasi Pendidikan Universitas Negeri Yogyakarta.* 18(1), 1-12.

Kartowagiran, Badrun. (2011a). Item and Test Analysis (Iteman). *Makalah Pelatihan Asesmen Pembelajaran Bagi Dosen Muda UPI*. Pascasarjana UNY. Yogyakarta.

_____. (2011b). Pengembangan Instrumen Asesmen Pembelajaran di Sekolah Bertaraf Internasional. *Makalah Penyusunan Bahan Ajar dalam Sertifikasi*. UIN Sunan Kalijaga. Yogyakarta.
_____. (2012). Penulisan Butir Soal. *Makalah Pelatihan Penulisan dan Analisis Butir Soal Bagi Sumber Daya PNS Dik-Rekinpeg*. Jakarta.
Kartowagiran, Badrun., Mardapi, Djemari., Subali, Bambang., Suyata, Pujiati., & Sriati, Arti. (1999). Peningkatan Kemampuan Menyusun dan Menganalisis Soal Tes Bagi Guru SMP Se D.I. Yogyakarta. *Jurnal Inoteks*. 1(1), 35-40.
Khoiriah. (2017). Prosiding Seminar Nasional Membangun Profesionalisme Guru Pendidikan Dasar Dalam Era Global. Jakarta.
Kubiszyn, Tom., & Borich, Gary. (2013). Educational Testing and Measuring: Classroom Application and Practice. *John Wiley and Sons, Inc.* The United State of America.
Kusairi, Sentot. (2012). A Computer-Assisted Analysis of Physics Formative Assessment For Senior High Schools. *Jurnal Penelitian dan Evaluasi Pendidikan.*69-87.
Lababa, Djunaidi. (2008). Analisis Butir Soal Dengan Teori Klasik: Sebuah Pengantar. *Jurnal Iqra.*5(1), 29-37.
Listyani, E., & Hidayati, K. (2010). Pengembangan Instrumen Kemandirian Belajar Mahasiswa. *Jurnal Penelitian dan Evaluasi Pendidikan.*
Mardapi, Djemari. (2003). *Bahan Lokakarya Metodologi Interaksi Pembelajaran.* Universitas Muhammadiyah Surakarta. Jawa Tengah.
Mardapi, Djemari., & Andayani, Sri. (2012). Performance Assessment dalam Perspektif Multiple Criteria Decision Making. *Prosiding Seminar Nasional Penelitian, Pendidikan dan penerapan MIPA UNY.*
Matondang, Zulkifli. (2009). Validitas dan Reliabilitas Suatu Instrumen Penelitian. *Jurnal Tabularasa Program Pascasarjana Unimed.*
Muljono, Pudji. (2007). Kegiatan Penilaian Buku Teks Pelajaran Pendidikan Dasar dan Menengah. *Artikel Staf Profesional BSNP Kegiatan Penilaian Buku Teks Pelajaran*.
Narayanan, Sowmya., & Adithan, M. (2015). Analysis of Question Papers in Engineering Courses With Respect to HOTS (Higher Order Thinking Skills). *American Journal of Engineering Education.*6(1), 1-10.
Nofiana, Mufida., Sajidan., & Karyanto, Puguh. (2016). Pengembangan Instrumen Evaluasi Higher Order Thinking Materi Kingdom Plantae. *Jurnal Pedagogi Hayati Program Studi Pendidikan Sains Pascasarjana Universitas Sebelas Maret.*
National Research Council. (1996). *National Science Education Standards.* National Academy Press. Washington, D.C.
_____. (2003). *National Science Education Standards*. National Academy Press. Washington, D.C.
Pappas, E., Pierrakos, O., & Nagel, R. (2012). Using Bloom's Taxonomy to Teach Sustainability in Multiple Contexts. *Journal of Cleaner Production.*
Popham, W. James. (1995). *Classroom Assessment: What Teacher Need to Know.* Allyn and Bacon. Los Angeles.
Rahayu, Septri., Akhsan, Hamdi., & Zulherman. (2014). Pengembangan Panduan Praktikum Perangkat Gelombang Mikro Materi Gelombang Elektromagnetik. *Jurnal Pendidikan*. 171-178.
Raisanen, Milla., Tuononen, Tarja., Postareff, Lissa., Hailikari, Telle., & Virtanen, Viivi. (2016). Students and Teacher Experiences of The Validity and Reliability of Assessment in a Bioscience Course. *Journal of Higher Education Studies*.6(4), 181-189.
Ramadhani, Martha Candra., Kantun, Sri., & Widodo, Joko. (2014). Analisis Validitas dan Tingkat Kesukaran Soal Latihan Evaluasi Akhir Tahun pada Buku Sekolah Elektronik (BSE) Mata Pelajaran Ekonomi SMA/MA Kelas XI. *Jurnal Artikel Hasil Penelitian Mahasiswa Universitas Jember.*1-7.
Ramos, J.L.S., Dolipas, B.B., & Villamor, B.B. (2013). Higher Order Thinking Skills and Academic Performance in Physics of College Students: A Regression Analysis. *International Journal of Innovative Interdisciplinary Research.*
Ratumanan, T.G., Laurens, T., & Mataheru, W. (2009). Pengembangan Model Pembelajaran Interaktif dengan Setting Kooperatif Model PISK. *Jurnal Matematika.*
Resnick, L.B. (1987). *Education and Learning to Think.* National Academy Press. Washington, DC.
Rofiah, Emi., Aminah, Nonoh Siti., & Ekawati, Elvin Yusliana. (2013). Penyusunan Instrumen Tes Kemampuan Berpikir Tingkat Tinggi Fisika Siswa SMP. *Jurnal Pendidikan Fisika.*1(2), 17-22.
Saido, Gulistan Mohammed., Siraj, Saedah., Nordin, Abu Bakar Bin., & Al Amedy, Omed Saadallah. (2015). Higher Order Thinking Skills Among Secondary School Students in Science Learning. *Journal of University Malaya.*3(4), 16-30.
Santrock, John W. (2014). *Psikologi Pendidikan Buku 2*. Edisi 3. Salemba Humanika. Jakarta.
Santyasa, I Wayan. (2005). *Analisis Butir dan Konsistensi Internal Tes.* Disajikan dalam Workshop Bagi Pengawas dan Kepala Sekolah Dasar Kabupaten Tabanan. Bali.
Sugiyono. (2011). *Metode Penelitian Kuantitatif , Kualitatif dan R&D.* Alfabeta. Bandung.

Sunarti & Rahmawati, Selly. (2014). *Penilaian dalam Kurikulum 2013.* Andi. Yogyakarta.
Suwarto. (2011). Teori Tes Klasik dan Teori Tes Modern. *Jurnal Widyatama*. Universitas Veteran Bangun Nusantara. Sukoharjo.
Uno, Hamzah B., dan Koni, Satria. 2014. *Assessment Pembelajaran.* Bumi Aksara. Jakarta.
Wardany, Kusuma., Sajidan., & Ramli, Murni. (2015). Penyusunan Instrumen Tes Higher Order Thinking Skill Materi Ekosistem SMA Kelas X. *Jurnal Biologi, Sains, Lingkungan dan Pembelajarannya.*SP-001-289:538-543.
Yahya, A.A., Toukal, Z., & Osman, A. (2012). Bloom's Taxonomy Based Classification for Item Bank Questions Using Support Vector Machines. *In Modern Advances in Intelligent System and Tools.* Springer. Berlin, Germany.
Yusuf, A. Muri. (2015). *Asesmen dan Evaluasi Pendidikan.* Prenadamedia Group. Jakarta.

